

Weekly Report

Done

1. 和彭老师, 巫老师一些讨论:

彭老师对于行为的 embedding 很感兴趣, 他们最近也做了一些工作, 尽管不是我们的*2vec, 但通过交流发现还是很像的. 他们本身也有想做*2vec 的想法. 他们将传统的按固定时间单位记录的人的行为信息, 转换为基于 Session 的人行为分析的方法. 这个 session 还可以进一步的聚合上去. 这里其实就带有两层 embedding.

那通过我们对于一些已有数据和网上常见可获取数据的分析, 我们觉得可以做事件序列的分析, 并且可以是带有层次的那种事件; 用 embedding 来浓缩的表示与展示信息, 用可视化探索一些常见的问题, 诸如异常检测, 关联分析, 用户肖像等.

那对于事件序列, 彭老师同时还提出了一种他们领域的, 将时变数据转换为网络数据, 他认为这个方法也可以做一些可视化.

结合彭老师近期一些论文, 近年来一些事件序列论文, 我将进一步思考发掘其中的点.

2. 一些琐事, 修改 PVis 论文, 与实验室其他同学的一些交流等等. 让小顾尝试一些 embedding 的基本方法和实现.

To Do

1. 阅读 Panpan 还有曹楠老师近期论文(事件相关的, 彭老师也觉得有意思). 结合彭老师的一些想法继续思考. 此外会和谢潇了解一下他们那边做 embedding 的一些思路, 看看是否可以借用.
2. 整理一下已有数据(以及那些在线的), 对于 KDD18 那篇 best paper (下面讲)上的方法论也要进行一些数据尝试.

论文阅读

1. **KDD18 Best Paper: Real-time Personalization using Embeddings for Search Ranking at Airbnb.** 本文来自 Airbnb, 在网络上有比较多的好评, 因为它比较可实践, 公式不复杂, 几乎是一篇公司工程实践的说明书. 其主体采用的仍然是 skip gram 模型, 主要区别于传统*2vec 的特点是: (1) 有一个 label 有知道他最后有没有订房间 作为 embedding 的额外信息 (2) 负采样的时候, 可能有同一个地区和不同地区的(我们住 airbnb 肯定尽量都一个地区); 我们要保证采一些相同地区的负样本, 否则有 bias (3) 推荐冷启动: 如果新用户咋给他实时推荐? 利用 metadata 来找类似的已有的 embedding 并去做平均 (4) 对于长时间的预测(一年旅行出差住个几次 airbnb, 区别于用户在网络上点击不同住房的实时预测) Listing (即 airbnb 住房) type 和 user type 嵌入到同一个空间里; 这两个 type 本身是对用户先做了聚合信息(因为次数太少, 单个人来看不能做 embedding), 再联合 embedding. 联合以后, 我们对一个人要做推荐, 只要找到这个人的 embedding 结果周围的 listing 即可推荐.

2. How to measure sessions of mobile phone use? Quantification, evaluation, and applications 彭老师的文章, 对于用户行为研究, 从按分钟计数不同行为, 转换为 session based, 其想要抓住这四个要素: (a) duration (amount of time), (b) frequency (number of tasks), (c) timing (start and end of each task), and (d) sequence (flow of adjacent tasks). 文中信息较多, 包括如何构建 session, 如何构建网络后进行 community detection 来对用户 session 进行聚类. 用户 session 自己聚类后, 怎么对用户-session cluster 之间进行聚类. 这样的方法也支持许多下游任务.
3. **KDD18** Embedding Temporal Network via Neighborhood Formation 研究的是动态网络边不断形成的一个过程, 进行建模并做 embedding. Embedding 还考虑了节点邻居间的相互影响. 这个思路比较新奇, 但我也想到以前确实见过研究网络生成过程的问题, 这个可能可以拿来做一些观察网络结构的问题, 也许东明可以去试试.
4. **Physical review letters** Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics 最初版本的从时间序列构建网络的方法, 据此还发现了一些彭老师原文中的公式错误....
5. 其余和同学讨论前后可能也过了一些论文, 不详述了.

工作时间

平时 30, 周末一共 8 小时. 总共 38 小时.